IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

Heart Disease Classification using PCA and Back-Propagation Neural Network

Shivani Sawai
shivani.sawai2912@gmail.com

Aanchal Koul
koulaanchalkoul@gmail.com

Antara Rangnekar
antara.rangnekar@gmail.com

Abstract –**Cardiovascular diseases are the leading cause of disability and premature death worldwide, and contribute substantially to the rising costs of health care. The fundamental anatomy pathological lesion is atherosclerosis, which occurs over the years and is usually advanced when symptoms appear, usually at maturity. Acute coronary and cerebrovascular events often occur suddenly and are often fatal before medical attention can be provided. It has been shown that the modification of risk factors reduces mortality and morbidity in people with cardiovascular diseases, diagnosed or not. The main objective of this research work is to develop a prototype which can determine and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. This paper uses Neural Networks Algorithm technique for heart disease prediction. PCA is used to reduce number of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. Performance of proposed approach is evaluated using confusion matrix plot.**

Keywords – **PCA, NN, Cardiovascular diseases etc.**

I. INTRODUCTION

Cardiovascular diseases (CVD) are defined as a group of conditions that affect the heart and blood vessels. Despite the decreasing trend shown in the last three decades in developed countries, at present, CVD as a whole are the main cause of mortality and hospitalization in the Indian population. The three basic cardiovascular problems, ischemic heart disease, cerebrovascular disease and heart failure, are based on their knowledge of the large epidemiological studies developed since the middle of the last century. The Framingham Heart Study [1], has become the most classic cohort study par excellence, establishing for more than six decades the essential role played by risk factors in the development of CVD. Primary cardiovascular prevention is based on the identification and control of cardiovascular risk factors in a healthy population, thus trying to prevent the onset of the disease [2].

In some developing countries, including India, in the last decades there has been a reduction in the mortality of CVD, especially of ischemic heart disease. However, there is a growing number of men and women who live with some CVD. The reasons for this apparent contradiction lie in the increase in life expectancy and improvements in the treatment and management of CVD.

CVD show differences between men and women. While in women, CVD represents the first cause of death, in men it is the second cause behind tumors. On the other hand, the trend of hospital morbidity rates of CVD in recent years has been a constant increase, both in men and women. In this sense, cardiovascular diseases are the first cause of hospitalization in the Spanish population. In the coming years an increase in the number of hospitalizations due to these diseases is expected, as a result of the technological development that will allow patients to offer new diagnostic and therapeutic instruments, the greater survival of patients with these health problems and the aging of the Indian population.

The three main CVD are, ischemic heart disease, cerebrovascular disease and heart failure, which together are responsible for 74% of mortality from vascular causes. In addition to this, ischemic heart disease and cerebrovascular disease constitute, respectively, the third and fourth cause of loss of adjusted life years due to disability.

Cardiovascular diseases are a public health problem of the first order and are expected to continue to be so in the future due mainly to population aging; in India, they represent the first cause of death and hospitalization. The three most important cardiovascular problems, ischemic heart disease, cerebrovascular disease and heart failure, are based on their knowledge of the major epidemiological studies, being The Framingham Heart Study [1], which has contributed the most to the development of cardiovascular prevention in the last six decades, providing solid evidence on the risk factors of these diseases.

The investigation of health results allows to evaluate the quality and effectiveness of health care, determined by obtaining pre-established final results. The measurement of such results is possible

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

through well-conceptualized indicators. Avoidable Hospitalization by Ambulatory Care Sensitive Conditions is a health indicator that, through the quantification of hospitalizations caused by a specific group of pathologies (including cardiovascular), aims to measure the resolutive capacity of primary care, based on the rationale that, the increase in preventive measures and the improvement of outpatient treatment at this level of care should correspond to a reduction in those hospitalizations.

II. Cardiovascular Diseases

Out of the 58 million deaths from all causes that are estimated to have occurred worldwide in 2005, cardiovascular diseases (CVD) represented 30%. This proportion is equal to that corresponding to the combination of infectious diseases, nutritional deficiencies and maternal and perinatal conditions [1]. It is important to note that a substantial proportion of these deaths (46%) were recorded in people under 70 years of age in the most productive period of life; Moreover, 79% of the morbidity burden attributed to cardiovascular diseases occurs in this age group [2].

Between 2006 and 2015, deaths due to non-communicable diseases (half of which correspond to cardiovascular diseases) are expected to increase by 17%, while deaths from infectious diseases, nutritional deficiencies and maternal conditions are estimated to be and combined perinatal will decrease by 3% [1]. Almost half the burden of disease in low- and middle-income countries is already due to non-communicable diseases [3].

A significant proportion of this morbidity and mortality could be prevented through population strategies and making cost-effective interventions accessible and affordable, both for people who already suffer them and for those who have a high risk of suffering them [3-5].

To address the growing burden of non-communicable diseases, in May 2000 the 53rd World Health Assembly adopted the WHO Global Strategy for the Prevention and Control of Non-communicable Diseases [6], and in doing so placed to non-communicable diseases within the global public health agenda. Since then, WHO has strengthened its efforts to promote primary prevention of non-communicable diseases throughout the population, through the Framework Convention on Tobacco Control [7] and the Global Strategy on Diet and Physical Activity [8]. These activities target common risk factors that are shared by cardiovascular diseases, cancer, diabetes and chronic respiratory diseases, and their implementation is essential to control the growing burden of non-communicable diseases. These measures should make it easier for healthy people to remain so, and those who suffer cardiovascular diseases or have a high cardiovascular risk change their behaviour. However, public health approaches throughout the population will not have a tangible immediate impact on cardiovascular morbidity and mortality and will only have a moderate absolute impact on the burden of disease [3, 4]. By themselves they cannot help the millions of people at high risk of cardiovascular disease (Table 1) or who already suffer from cardiovascular disease. A combination of population-wide strategies and strategies targeting people at high risk is needed to reduce the burden of cardiovascular disease. The extent to which one strategy should have a preponderance over another will depend on the actual effectiveness that can be achieved, as well as its cost-effectiveness and the availability of resources [1-4].

Although cardiovascular disease already poses a considerable economic burden for low and middle income countries [9], the resources available for their management in these countries are limited, since there are competing health priorities. However, it is essential to recognize that the transition to lower levels of infectious diseases and to higher levels of non-communicable diseases is already underway. If action is not taken now, there will be a large increase in preventable cardiovascular diseases, which will be an important pressure for national economies [10-12]. In this context, it is imperative to focus limited resources on those most likely to benefit from them. Therefore, as foreseen in the Global Strategy for the Prevention and Control of Non-communicable Diseases [6], one of the main tasks of WHO and its Member States is to establish, on a larger scale, cost-effective and integrated approaches to the prevention of cardiovascular diseases.

III. Control of Cardiovascular Diseases

In all populations it is essential that the approach aimed at the high-risk population be complemented by public health strategies throughout the population (Figure 1) [11]. Although cardiovascular events are less likely to occur in people with low risk levels, there is no level of risk that can be considered "safe" [13]. Without preventive public health efforts in the entire population, episodes of cardiovascular disease will continue to affect people with low and moderate risk levels, which are the majority in any population. In addition, public health approaches can effectively arrest the development of atherosclerosis (and also reduce the incidence of some cancers and chronic

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

respiratory diseases) in young people, thereby decreasing the likelihood of future epidemics of cardiovascular disease, as seen in the period 1960-1990 in most countries with high incomes. The strategies oriented to the whole population will also support the modification of the way of life in people at high risk. The degree to which one measure should be emphasized with respect to another depends on the actual effectiveness that can be achieved, its cost-effectiveness, as well as considerations regarding resources.

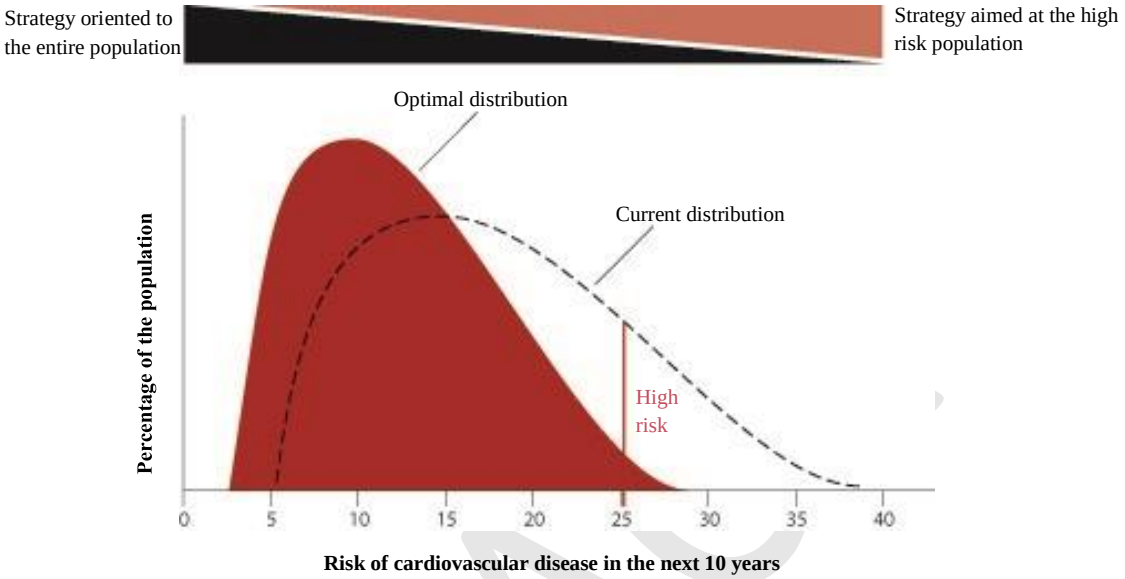

Figure 1: A combination of strategies targeting the entire population and the high-risk population is needed to reduce the distribution of cardiovascular disease risk in the population (so that the distribution of cardiovascular risk shifts to the left)

IV. Proposed Methodology

The problem with risk factors related to heart disease is that there are many risk factors involved like age, usage of cigarette, blood cholesterol, person's fitness, blood pressure, stress and etc. and understanding and categorizing each one according to its importance is a difficult task. Also a heart disease is often detected when a patient reaches advanced stage of the disease. Hence the risk factors are analyzed from various sources [14]-[15]. The dataset was composed of 12 important risk factors which were sex, age, family history blood pressure, Smoking Habit, alcohol consumption, physical inactivity, diabetes, blood cholesterol, poor diet, obesity .The system indicated whether the patient had risk of heart disease or not. The data for 50 people was collected from surveys done by the American Heart Association [15]. Most of the heart disease patients had many similarities in the risk factors [16]. Table 1 shows the identified important risk factors and the corresponding values and their encoded values in brackets, which were used as input to the system.

Table 1: Risk factors values and their encodings [17]

S. No.	Risk Factors	Values
1	Sex	Male (1), Female (0)
2	Age (years)	20-34 (-2), 35-50 (-1), 51-60 (0), 61-79 (1) , >79 (2)
3	Blood Cholesterol	Below 200 mg/dL - Low (-1) 200-239 mg/dL - Normal (0) 240 mg/dL and above - High (1)
4	Blood Pressure	Below 120 mm Hg- Low (-1) 120 to 139 mm Hg- Normal (0) Above 139 mm Hg- High (1)
5	Hereditary	Family Member diagnosed with HD -Yes (1) Otherwise –No (0)
6	Smoking	Yes (1) or No (0)
7	Alcohol Intake	Yes (1) or No (0)
8	Physical Activity	Low (-1) , Normal (0) or High (1)
9	Diabetes	Yes (1) or No (0)
10	Diet	Poor (-1), Normal (0) or Good (1)
11	Obesity	Yes (1) or No (0)
12	Stress	Yes (1) or No (0)
Output	Heart Disease	Yes (1) or No (0)

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

Data analysis has been carried out in order to transform data into useful form, for this the values were encoded mostly between a range [-1, 1]. Data analysis also removed the inconsistency and anomalies in the data. This was needed. Data analysis was needed for correct data pre-processing. The removal of missing and incorrect inputs will help the neural network to generalize well. Moreover the principal component analysis is used for attribute minimization.

Principal Component Analysis

PCA is a linear technique used for the elimination of data redundancy. It is widely used, however, its greatest limitation is based on the assumption of linearity.

PCA allows a base change to one of smaller dimensionality of data consisting of a large no. of attributes X through the transformation equation $Y = PX$, where P is an orthogonal matrix called the representation matrix. The objective is to determine the matrix P that allows the data cloud to be projected to a space of smaller dimension.

The strategy is to look for P so that the non-correlation between vectors of Y is guaranteed, that is, $C_{ij} \in C_Y, i \neq j$ is null. If the correlation between the different samples is zero, the redundancy is eliminated and the data subspace can be described by P. Otherwise, each C_{ij} entry corresponding to large values that will represent high redundancy of the observations i and j and, therefore, there will be the present noise. The algorithm to find P starts with the centering and standardized data. Then, we compute the covariance matrix of X, $C_X = \frac{1}{n} XX^T$ that is symmetric and diagonalizable, and that quantifies the covariance between the measurements. Then, we obtain the eigenvectors of C_X, which are chosen as vector columns of P, ordered according to the eigenvalue and which serve as new coordinates of the system where the variance is maximized. The appropriate number of eigenvectors are chosen, which are called principal components and which describe the information of the data set according to their coefficient of inertia, which indicates the percentage of this, present in each principal component.

Steps for performing a principal component analysis [18]:

1. Take the entire dataset containing D-dimensional samples.
2. Calculate the D-dimensional mean vector (i.e., the mean for every dimension of the whole dataset).
3. Calculate the covariance matrix of the entire dataset.
4. Calculate the eigenvectors $(E_1, E_2, \ldots, E_D)$ and corresponding eigen values $(\gamma_1, \gamma_2, \ldots \gamma_D)$.
5. Categorise the eigenvectors by reducing eigen values and select n eigenvectors with the largest eigen values to form a $D \times n$ dimensional matrix M (where every column represents an eigenvector).
6. Utilize this $D \times n$ eigenvector matrix to convert the samples onto the new subspace. This can be summarized by the mathematical equation: $Y = M^T \times x$ (where x is a $D \times 1$-dimensional vector representing one sample, and Y is the transformed $n \times 1$-dimensional sample in the new subspace)
7. Find standard deviation using the following formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (1)$$

In this paper, Neural Network algorithm based approach is used to determine optimum number of clusters in analyzed data.

Neural Networks

Artificial neural networks allowed us to formally simulate the work of the human brain, scientists have almost discovered how the human brain's work in terms of scalability and portability of learning memory and the ability to distinguish objects and the ability to make decisions and as we know, the brain is made up of billions of neurons interconnected in a very complex way by neuronal cells, forming an enormous network of neurons associated with them to each other.

This correlation between nerve cells gives them the ability to store and deliver information, images, audio and signal sequences that receive across different neurons, neural networks also allow learning through repetition and error.

Behaviour of Artificial Neuron

Neuron inputs "X": These come either from other "processors" (neurons) or from the environment.

The weight "W" (synaptic coefficient): Is a numerical value associated with a connection between two units (neurons) that reflects the strength of relationship (connection) between these two units i and j, and it is noted by W_i.

The Aggregation Function (combination) "P": It combines inputs and weights by calculating the influence of each entry taking into account its

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

weight. This influence is calculated via the next formula:$P = \sum W_i X_i$, where W_i is the weight of the connection at input i, X_f: is the signal of the input i.

Transfer Function (activation): The activation function (the transfer function) plays a very important role in the behaviour of the neuron. It returns a value representative of the activation of the neuron, this function has as parameter the weighted sum of the entries as well as the threshold of activation. It calculates the output value from the result of the combination function: $S = F(P)$.

Where, S: is the output value, F: is the transfer function.

Multilayer Perceptron (MLP) of Neural Network

The multilayer perceptron is an organized network of artificial neurons organized in layers and where the information travels in one direction, from the input layer to the output layer.

The implementation of the neural networks includes both a design part, the objective of which is to make it possible to choose the best possible architecture, and a part of numerical calculation, to realize the learning of a neural network, but in order to improve the functioning of the MLP on one side and to reduce the computation time as much as possible, one must look for an optimal architecture in terms of number of layers, number of neurons per layer and number of possible outputs.

From a given neural network architecture and available examples (the learning base), the optimal weights are determined by the algorithm of the back propagation of errors so that the output of the model approaches as much as possible of the desired operation [19].

Learning MLP networks: The learning of a multilayer neural network is generally done by the back propagation algorithm, which is the most widely used supervised learning example for MLPs. It uses an optimization method. Universal is to find the network coefficients (weight) minimizing a global error function (cost function).

The gradient back propagation is a method for calculating the gradient of the error for each neuron in the network, from the last layer to the first, the principle of back propagation can be described in three basic steps:

Back Propagation: After calculating the learning error which is the result of step of forward propagation of the inputs, in this phase will back-propagate this error through the network layers going from the outputs to the inputs (towards the back), this error will thus be distributed to the neurons of hidden layers in order to be able to adjust in the next phase the weights of the network, the computation of the error (Δ) of each neuron j of the hidden layer is made using the following formula:

$$\Delta_j = VA_j(1 - VA_j)\sum_{k \in succ(j)} W_j K \Delta_k \quad (2)$$

The Update of the Weights: At the end of the previous step, the learning error was distributed (Back Propagation) to all the neurons of the hidden layers, and now in the current phase we recalculate the weights of the network which are previously initialized with random values using the following rule:

New weight = old weight + learning rate * current neuron error * output of the previous layer. This is described by the formula:

$$W_{ij} = W_{ij} + \alpha * \Delta_j * VA_i \quad (3)$$

Where, α is the learning rate, which is usually in the range [0,1] (chosen by the user).

During the learning of the neural network, these three phases (forward propagation, back-propagation and updating of weights) are repeated as many times as the number of examples of the learning base. Once completed, the mean squared error (MSE) is computed, this measure is often applicable in binary classification where the classes are 0 and 1. This is a measure related to probabilities, it is defined by the formula:

$$MSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(d_k - y_k)^2} \quad (4)$$

This is a network validation step where:

d_k: the desired value in the learning example.

y_k: the neuron value of the output layer calculated by the MLP.

The obtained MSE value must be below a certain threshold for which the model obtained can be said to respond well to the learning examples, otherwise the procedure will be repeated with other initial values of the W_{ij} weights and the learning rate α.

The learning phase of a neuron network can therefore be summarized by the following algorithm:

Initialization of weights with random values included in a chosen interval
Reading learning examples.
Normalize the training data;
Repeat
For *each learning example* ***Do***
    Propagation of the entrance to the front
    Propagation of the error towards the back (back-propagation)
    Update weights
End for
Computation of the MSE
As long as *the MSE is greater than the threshold or maximum number of iterations*

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

V. Simulation And Results

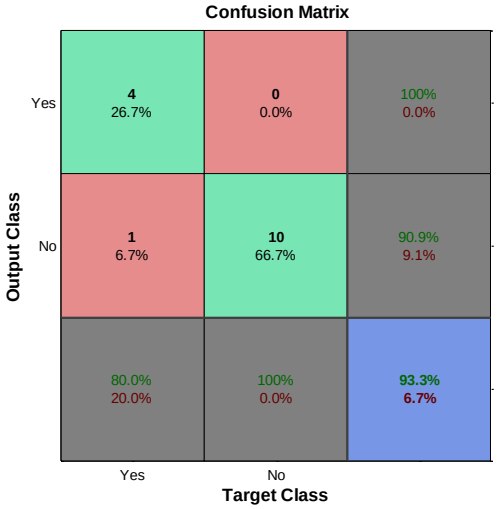

Figure 2: confusion matrix for heart diseases detection

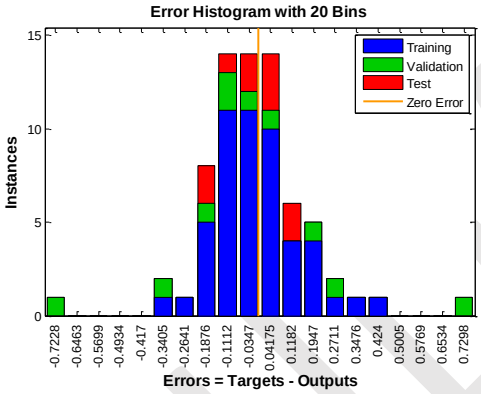

Figure 3: Histogram of training and testing

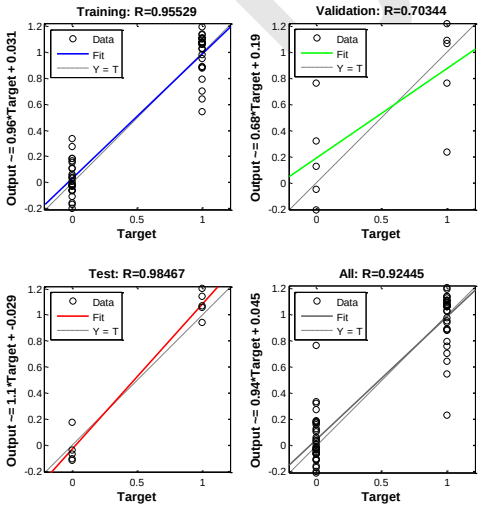

Figure 4: Regression plot of neural network

VI. Conclusion

A serious epidemiological problem in the contemporary world constitutes cardiovascular diseases. In the future, cardiovascular prevention will be the field of research development and the destination of human and economic resources, aimed at controlling an expansion that does not seem to stop until today. A range of conditions allow to predict the existence of serious clinical episodes, are the risk factors, which compromise the future of the patient. Efforts have been continued to achieve knowledge of the main cardiovascular risk factors and thus expand the existing knowledge on clinical, pathophysiological, epidemiological or therapeutic aspects of cardiovascular diseases, The World Health Organization (WHO) and its Department of Chronic Diseases and Health Promotion have developed the general framework for the prevention and control of chronic diseases. The strategic objectives are to raise awareness about the epidemic of chronic diseases; create healthy environments, especially for poor and disadvantaged populations; curb and reverse the tendency to increase the common risk factors of chronic diseases, such as unhealthy diet and physical inactivity; and prevent premature deaths and avoidable disabilities caused by major chronic diseases.

Various algorithms are exist in literature for the prediction of heart disease using major risk factors. This paper uses heuristic techniques to provide the input to our network to give better results and it was found that it is effective to predict the risk of heart disease when the person provide the required attributes value. The confusion matrices in simulation results show that the proposed PCA-NN based method outperforms the Neural Network based approach with the accuracy of 93.3 %.

Reference

[1] World Health Organization. "Preventing chronic diseases a vital investment." In Preventing chronic diseases a vital investment. 2005.

[2] World Health Organization. The world health report 2002: reducing risks, promoting healthy life. World Health Organization, 2002.

[3] Lopez, Alan D., Colin D. Mathers, Majid Ezzati, Dean T. Jamison, and Christopher JL Murray. "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data." The Lancet 367, no. 9524 (2006): 1747-1757.

[4] Manuel, Douglas G., Jenny Lim, Peter Tanuseputro, Geoffrey M. Anderson, David A. Alter, Andreas Laupacis, and Cameron A. Mustard. "Preventive medicine: Revisiting Rose: strategies for reducing coronary heart disease." BMJ: British Medical Journal 332, no. 7542 (2006): 659.

[5] World Health Organization. Prevention of recurrent heart attacks and strokes in low and middle income populations: evidence-based recommendations for

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 10, May 2018)

policy-makers and health professionals. World Health Organization, 2003.
[6] World Health Organization. "Global action plan for the prevention and control of noncommunicable diseases 2013-2020." (2013).
[7] FCTC, WHO. "WHO Framework Convention on Tobacco Control. Geneva." (2003).
[8] World Health Organization. "Global strategy on diet, physical activity and health." (2004).
[9] Leeder, Stephen, Susan Raymond, Henry Greenberg, Hui Liu, and Kathy Esson. "A race against time: the challenge of cardiovascular disease in developing economies." New York: Columbia University (2004).
[10] Miranda, J. Jaime, Sanjay Kinra, Juan P. Casas, G. Davey Smith, and Shah Ebrahim. "Non-communicable diseases in low- and middle- income countries: context, determinants and health policy." Tropical Medicine & International Health 13, no. 10 (2008): 1225-1234.
[11] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of IEEE Conference on Information and Communication Technologies (ICT), pp. 1227–1231, April 2013.
[12] World Health Organization, and World Health Organization. Cardiovascular Disease Programme. WHO CVD-risk management package for low-and medium-resource settings. World Health Organization, 2002.
[13] Jackson, Rod, John Lynch, and Sam Harper. "Preventing coronary heart disease: Does Rose's population prevention axiom still apply in the 21st century?." BMJ: British Medical Journal 332, no. 7542 (2006): 617.
[14] Centre for Disease Control and Prevention, Online available at: http://www.cdc.gov/heartdisease/risk_factors.htm
[15] American Heart Association, Online available at: http://www.heart.org/HEARTORG/Conditions
[16] D. Isern, D. Sanchez, and A. Moreno, "Agents Applied in Health Care: A Review", International Journal of Medical Informatics, Vol. 79, Issue 3, pp. 146-166, 2010.
[17] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of IEEE Conference on Information and Communication Technologies (ICT), pp. 1227–1231, April 2013.
[18] Jabbar, M. Akhil, B. L. Deekshatulu, and Priti Chandra. "Classification of heart disease using artificial neural network and feature subset selection." Global Journal of Computer Science and Technology Neural & Artificial Intelligence 13, no. 3 (2013).
[19] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." Atmospheric environment 32, no. 14-15 (1998): 2627-2636.